DOCUMENT RESUME

ED 441 027 TM 030 825

AUTHOR Finney, Sara J.; Smith, Russell W.; Wise, Steven L.
TITLE The Effects of Judgment-Based Stratum Classifications on the Efficiency of Stratum Scored CATs.
PUB DATE 1999-04-00
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Classification; *Computer Assisted Testing; *High School Students; High Schools; Item Banks; Mathematics Tests; *Test Construction; Test Items
IDENTIFIERS Stratification

ABSTRACT

Two operational item pools were used to investigate the performance of stratum computerized adaptive tests (CATs) when items were assigned to strata based on empirical estimates of item difficulty or human judgments of item difficulty. Items from the first data set consisted of 54 5-option multiple choice items from a form of the ACT mathematics assessment. Items from the second data set were drawn from a computerized algebra test with sample sizes of 250 to 300 examinees for each of the 140 options. Each of the 11 judges independently sorted items into difficulty strata for one of the two datasets. It was found that stratum CATs based on empirical item difficulties (both p-values and b-parameters) have increased efficiency and precision relative to a conventional fixed-length test. It was also shown that efficiency and precision increased as the number of strata increased. Also, under certain conditions, the stratum CAT was able to match or exceed the precision and efficiency of the traditional CAT. These findings provide evidence of the promise stratum CATs have as a non-Item Response Theory adaptive testing method that requires minimal item data. The results from the stratum CAT based on human judgment were not promising. Findings suggest that stratum CATs based on human judgments do not provide increased efficiency or precision over a conventional fixed-length test. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 027

The Effects of Judgment-Based Stratum Classifications on the Efficiency of Stratum Scored CATs

Sara J. Finney & Russell W. Smith

University of Nebraska-Lincoln

Steven L. Wise

James Madison University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

S. Finney

TM030825

Paper presented at the 1999 annual meeting of the National Council on Measurement in Education, Montreal, Canada.

BEST COPY AVAILABLE

The Effects of Judgment-Based Stratum Classifications on the Efficiency of Stratum Scored CATs

Computerized adaptive testing has become an attractive tool for examiners due to the efficiency it provides relative to the traditional fixed length test. The high efficiency is accomplished by continually estimating an examinee's proficiency and administering items that best match this estimated proficiency. Currently, virtually all computerized adaptive tests (CATs) use item response theory (IRT) for item calibration, item selection and proficiency estimation.

While many large-scale testing programs have taken advantage of the efficiency CATs can provide, they are rarely used in testing situations that are characterized by small numbers of examinees (e.g. classroom testings). Wise (1999) listed several obstacles to CATs being implemented in classroom settings. These obstacles centered around the issue of employing IRT methods to implement the adaptive test. Specifically, teachers in small classroom settings may lack an item pool with an adequate number of examinee responses to calibrate the items. In addition, they may lack both the psychometric skill needed to implement a CAT and the software used to calibrate items and administer the tests adaptively.

To address these problems, Wise (1999) developed a new method of adaptive testing, the stratum CAT, which does not require the use of IRT. The stratum CAT is based on an adaptive administration method developed by Weiss (1973) called stradaptive testing. Items in a stradaptive test are initially sorted into ranked strata based on item difficulty. The selection of an item to be administered is determined by both examinee performance and the item's stratum membership. For example, an examinee beginning a test is administered an item from a middle stratum. If the item is passed, an item from the next higher stratum is administered; if the item is failed an item from the next lower

stratum is administered. This one-up, one-down process continues until a specified number of items have been administered. While the administration of items does not rely on IRT methods, stradaptive testing does employ IRT for the assignment of items to strata, selection of items from strata, and the estimation of examinee proficiency. Specifically, Weiss (1973) used IRT difficulty and discrimination parameters to sort items into ranked strata and used the average of the IRT difficulty parameters from administered items as an estimate of examinee proficiency. Because a CAT without reliance on IRT was desired by Wise (1999), the stratum CAT uses the same one-up one-down administration method used in the stradaptive test but does not use the IRT based methods for sorting items, selecting items from strata, or scoring.

The scoring method developed to estimate proficiency for a stratum CAT is called stratum scoring (Wise, 1999). Items are rank ordered by difficulty and subdivided into a predetermined number of strata. The strata are then assigned unit weights that are used for scoring purposes. These weights were designed so that correct responses to more difficult items are given more credit than correct responses to less difficult items. Similarly, incorrect responses to difficult items are penalized less than incorrect responses to less difficult items. For example, if an item pool was broken down into 3 strata, correct responses to items in stratum 1, 2 and 3 would be awarded 1, 2 and 3 points, respectively. Similarly, incorrect responses to items in stratum 1, 2 and 3 would be assigned –3, -2 and –1 points, respectively. After administering the desired number of test items, an examinee's item scores would be summed to compute the total stratum score. This score then represents the examinee's proficiency as measured by the test.

Wise (1999) completed two simulation studies to evaluate the difference between the precision of estimated proficiency for stratum CATs, conventional linear tests, and traditional CATs using IRT. Both studies investigated the effects of using different

numbers of strata, different length tests, and different item pools (one-parameter and three-parameter) on the recovery of true proficiency. The major differences between the two studies were the sorting method used to assign items to strata and the stability of the difficulty estimates used for the sorting.

The first study investigated the proficiency of the stratum CAT when items were assigned to strata using generated b-parameters that were assumed to be the true difficulty parameters of the items. The second simulation study examined the performance of the stratum CAT when item difficulty was estimated using classical test theory from small samples of examinees (n=50 & 100). Both studies showed that the stratum CAT produced scores that were more precise than the conventional linear test and the precision increased as the number of strata increased.

Furthermore, the stratum CAT seems to be robust to item misclassification. The stratum CAT continued to outperform the conventional fixed length test even though several items were misclassified when p-values were used to assign items to strata. In addition, the 9-stratum CAT based on p-values from a three parameter item pool exceeded the precision of the traditional CAT. These results suggest that adaptive tests using classical test theory are more efficient than conventional linear tests and can possibly match or exceed the efficiency of a traditional CAT when item parameters are derived from small samples.

The results from the stratum CAT using classical test theory for assignment of items to strata are encouraging. Since classroom teachers usually have experience with classical test theory and typically have a small number of examinees, this method of adaptive testing could provide them with a more efficient testing experience without the use of IRT. Teachers would simply need to calculate each item's p-value and then subdivide the set of items into the desired number of strata.

Wise (1999) found that the stratum CAT can perform well when item strata are based on p-values calculated from a small number of examinees. However, there may be situations in which examiners wish to use the stratum CAT but item responses from even 50 examinees are not available to estimate item difficulty. Another possible source of data that could be used to sort items into strata is human judgment of item difficulty. Wise's finding that the stratum CAT outperformed the conventional fixed length test even though numerous items were misclassified is encouraging. This finding suggests that the stratum CAT can tolerate some misclassification of items into strata. The purpose of this study is to investigate if human judgments can be used to sort items into strata well enough to benefit from the efficiency of the stratum CAT relative to a conventional fixed length test.

Early research on human judgment of item difficulty was conducted by Lorge and Kurglov (1952, 1953). They compared judges' abilities to estimate the relative and absolute difficulty of arithmetic items. Absolute difficulty refers to an estimate of the percentage of examinees passing an item, while relative difficulty refers to the rank order of the items by difficulty. The judges consisted of advanced students in test construction for the first study, while the judges in the second study were experienced teachers of mathematics. They found that judges in both studies were able to estimate relative item difficulty well but could not estimate absolute difficulty. The range in correlations between individual estimated difficulty rankings and empirical difficulty rankings showed the variability in subjective judgments of relative item difficulty (.52 to .84). They also found that the pooled estimates of difficulty rankings across the judges provided better estimates for difficulty than individual estimates. The effect of providing information regarding the empirical difficulty of several anchor items prior to beginning the estimation of item difficulty was also examined. They found that

having prior item difficulty information makes no appreciable difference in the estimates of relative item difficulty.

Quereshi and Fisher (1977) investigated the ability of 5 judges to estimate the relative item difficulty of 44 letter series items. The judges were selected due to their practical experience administering and interpreting psychological tests. As Lorge and Kurglov (1952, 1953) found, the judges were able to estimate item difficulty rank well, but again there was a great deal of variation in their ability. The correlations between estimated and empirical relative difficulty ranged from .44 to .62. Also, the best estimate of relative item difficulty was from the pooled judgments across the 5 judges.

Green (1983) used the method of paired comparisons to examine the ability of 19 judges to discriminate between the difficulties of 10 astronomy items. This method consisted of participants identifying the more difficult item of each of the 45 item pairs. For each participant, the item rankings were then derived from these judgments and compared to the empirical rankings using Goodman and Kruskal's gamma. The mean value of gamma was .20 which suggests that participants are not able to judge relative item difficulty well.

Wise, Finney, Enders, Freeman & Severance (in press) also found that judges are not highly proficient at discriminating item difficulty. Undergraduate students were asked to identify the more difficult item for each of 30 item pairs of ACT mathematics items. The percentage of correct difficulty judgments across judges was only slightly above 50%, which would be the expected percentage under random guessing. A second group of students completed the task but in addition they were required to solve each item. While the percentage of correct difficulty judgments across the judges increased (65%), difficulty judgment performance remained poor.

The ability of judges to estimate relative item difficulty is not clear cut. Also, the extent to which items need to be ranked correctly in order to benefit from the efficiency of the stratum CAT is unknown. If judges are able to estimate relative item difficulty fairly well, empirical estimates of item difficulty may not be needed to effectively groups items for a stratum CAT. This study investigated the performance of a stratum CAT when the assignment of items to strata is based on human judgment of difficulty. Some previous studies in this area have purposefully selected judges based on knowledge in the area of measurement or the content area of the items (Lorge & Kurglov,1952, 1953; Quereshi & Fisher, 1977) . The current study used judges of both types to investigate any difference between the types of raters in the consequent performance of the strata CAT.

In addition to investigating the performance of the stratum CAT based on judged item difficulty, extensions of the two previous simulation studies of stratum CATs (Wise 1999) were also examined. Previously, generated data sets were used to examine the efficiency of the stratum CAT. The current study used items from two operational tests to compare the efficiency of stratum CATs when empirical estimates of item difficulty or human judgments of item difficulty are used for sorting items into strata. The two item pools were selected on the basis of their difference in content, numbers of items and IRT model used to calibrate the items. It was expected that that stratum CATs using item sortings based on empirical difficulty estimates would yield scores that were more precise than scores from stratum CATs using item sortings based on human judgments. It was also expected that a stratum CAT based on combined rankings from the judges would perform better than a stratum CAT based on individual judgments. No specific hypotheses were made regarding the differences in

the performance of stratum CATs due to the specific group of judges completing the sorting.

Method

Test Materials

Test items from two empirical datasets were used in this study. Items from the first dataset consisted of 54 five-option multiple choice items from a 60-item form of the ACT Mathematics test. The last 6 items from this 60-item form were not used in the current study due to potentially unstable difficulty estimates caused by their position on this timed test. Item difficulty parameters were obtained by fitting the three parameter IRT model to the item responses from a sample of 1000 high school students who had previously taken the ACT Mathematics test. Each empirically estimated difficulty parameter represented an item's "true" difficulty for the purpose of this study. Item difficulty ranged from -.8 to 2.0.

Items from the second dataset were drawn from a computerized algebra test designed to assess whether students possess the algebra skills necessary to be successful in an introductory statistics course. All 140 four-option multiple choice test items from this item pool were used in the current study. The pool was calibrated using a modified one-parameter IRT model that specified a 0.20 common lower asymptote. Sample sizes of 250 to 300 examinee responses per item were available. Again, the empirically estimated item difficulty based on the calibration sample represented the item's true difficulty for the purpose of this study. Item difficulty ranged from –5.3 to 4.0.

Each item was placed on a four by six inch index card for judging purposes. Care was taken to replicate the appearance of the item on the original test as closely as possible. A unique random three-digit item identification number was assigned to each item and placed in the lower right hand corner of each card. For each item, an asterisk

was placed next to the correct option. This was done to better represent the item information that would be available to teachers involved in the process of judging item difficulty.

Participants

The eleven judges were purposefully chosen for the study based on experience in the area of mathematics or measurement. Specifically, four judges (one professor and three graduate students) from a university Mathematics department and four judges (two professors and two graduate students) from a university Educational Psychology department ranked the difficulty of the 54 ACT Mathematics items. Due to the small number of professors and graduate students in each department and the length of time needed to rank the 140 algebra items, only three judges (one professor and two graduate students) from the Educational Psychology department participated in the ranking of these items. All seven judges from the Educational Psychology program met the following criteria: (a) previous experience instructing a statistical methods course, (b) practical experience in the development, administration and interpretation of tests and (c) completion of at least two years of study in a graduate program in Educational Psychology. The four judges from the Mathematics department met the following criteria: (a) previous experience instructing a mathematics course (b) completion of at least two years of study in a graduate program in Mathematics.

Procedure

Each of the eleven judges independently sorted items into difficulty strata for one of the two datasets. A shuffled stack of the 54 ACT items, a Recording Sheet, and the following written instructions were placed in front of each of the 8 judges selected to sort the ACT items prior to beginning the sorting:

You have been given a stack of 54 ACT math items. Your task is to sort these items into ordered groups, or strata, based on the items' difficulty. You should consider difficulty in terms of the group of high school students taking the ACT. You will sort the items twice, first into 3 strata then into 5 strata.

Task 1: Sorting into 3 Strata

The 54 items are to be sorted into strata based on item difficulty: items belonging to stratum 1 being the easiest and items belonging to stratum 3 being the most difficult. For each item, the correct option is identified by an asterisk. Upon completing the sorting, each of the 3 strata should have 18 items. After you have completed sorting the items, record the identification numbers of the items belonging to each stratum on the Recording Sheet. In the bottom right-hand corner of each item is a three digit random number for item identification purposes only. On the Recording Sheet, record the identification number for each item under the appropriate stratum number. There is no need to order the items within each stratum. After completing this task, notify the researcher.

When the judges completed sorting the items into 3 strata, they were given 5 manila envelopes labeled 1 to 5 and given the following instructions:

Task 2: Sorting into 5 Strata

For the second task, the same items are to be sorted into 5 strata: items in stratum 1 being easiest and items in stratum 5 being the most difficult. When sorting the items, group the following number of items in each stratum:

Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
13	9	9	9	14

For this task, instead of recording identification numbers, you are to place the items belonging to each stratum into the envelope labeled with the stratum number. At this point you have completed the task.

The number of items assigned to each stratum was determined based on the desired length of the simulated tests. Test lengths of 20 and 15 items were planned for the 3-strata sorting which could be accomplished with 18 items belonging to each stratum. A 15-item test was planned for the 7-strata sorting which prevented simply assigning an equal number of items to each stratum. Since high (low) proficiency examinees can potentially stay at the highest (lowest) strata by consistently passing (failing) these items, additional items were assigned to these strata to guarantee enough items for the

desired test length. The ranking procedure for the ACT items was completed by the judges in approximately 1 hour.

The 3 judges ranking the 140 algebra items received similar directions. The only changes made to the procedure were that items were to be first sorted into 5 strata of 28 items each. After completing the 5-strata sorting, they were to sort the 140 items into 7 strata with 25 items placed in strata 1 and 7 and with 18 placed in all other strata. Again, the number of items assigned to each stratum was determined by the desired test lengths, which were 15, 20, 25 and 30 items. The ranking procedure for the algebra items was completed in approximately 1.5 hours.

Computer Program

Fortran 77 computer programs developed by Wise (1999) were used to simulate a conventional test, a traditional CAT and stratum CATs for each of the two item pools. Specifically, the ACT items were used to simulate 3-strata and 5-strata CATs in which the items were sorted into strata based on p-values and human judgment. The algebra items were used to simulate 5-strata and 7-strata CATs in which the items were sorted into strata based on b-parameters and human judgment. The b-parameters were used instead of the p-values because the original item calibration data were not available for calculating p-values. Since the algebra items were calibrated using a modified one-parameter model, the rank order of items based on b-parameters should be identical to the rank order of items based on p-values. Therefore, the classification of items into strata based on b-parameters is a meaningful substitute for the non-IRT classification of items based on p-values.

For each test condition, true proficiency values for 10,000 hypothetical examinees were randomly generated from a standard normal distribution. The empirical item parameters were used in all three testing conditions to calculate the probability of

passing each item. Therefore, the three-parameter model was used for the ACT items while a modified one-parameter model was used for the algebra items. For each item a uniform random number between 0 and 1 was generated. If the probability of passing the item by a given examinee exceeded the random number then the item was passed; otherwise, it was failed.

The conventional test was conducted by simply simulating the administration of the number of items of a given test length. The test began by randomly selecting an initial low-difficulty item and then choosing every nth item from the difficulty ranked set of items until the test length had been reached. This resulted in a broad ranged conventional test. For the 54 ACT items, the 15-item test began by randomly selecting an initial easy item and then choosing every 3rd item until 15 items had been administered. For the 140 algebra items, the same procedure was followed except every 9th item was selected. The proficiency estimate for each hypothetical examinee was the number of items passed. This same procedure was used for every conventional test. In a given condition, the same conventional test was used for all hypothetical examinees.

The traditional CAT simulated a fixed-length adaptive test based on the three-parameter model for the ACT items and the modified one-parameter for the algebra items. An initial proficiency estimate of 0 was used, and a maximum information criterion was used to select the item to be administered at each step of the CAT. The proficiency estimate was bounded at –5.0 and +5.0 to prevent nonconverging proficiency estimates.

The stratum CAT programs simulated fixed-length adaptive tests using the stradaptive method. The first item administered was randomly selected from the middle stratum. If the item was passed an item from the next higher stratum was administered; if an item was failed an item from the next lower stratum was

administered. If an item from the highest stratum was passed, the examinee continued to receive items from this stratum until one was failed. The same procedure was followed for examinees failing items in the lowest stratum. The sum of the item stratum scores represented the proficiency estimate for the examinee. This procedure was used for the stratum CATs based both on human judgment of item difficulty and empirical estimates of item difficulty.

Results and Discussion

ACT Items

The correlations between the true and judged item strata membership for the ACT items are presented in the top portion of Table 1. For both groups of judges, the ranges of correlations across both 3 and 5 strata were lower than those reported in previous studies (Lorge & Kurglov,1952, 1953; Quereshi & Fisher, 1977). Also, the correlations tended to decrease as the number of strata increased. Understandably, judges were better able to make crude judgments about item difficulty than more precise judgements. The correlations based on the pooled judgments were higher than the correlations based on individual judgements for only the Mathematics judges. Across both numbers of strata, Educational Psychology Rater A sorted items more accurately than the sorting produced by pooling the judgements of the Educational Psychology raters.

The results of the simulations using the ACT items are presented in Table 2. The traditional CATs using maximum likelihood estimation (MLE) did not perform well. Approximately 8% of the hypothetical examinees at each test length were assigned –5 or +5, the boundary MLE estimates. This is most likely due to the small range of difficulty (-.8 to 2) for these test items and the small number of items administered. Since for

many of the examinees an adequate estimate of proficiency could not be calculated using MLE with 20 items or less, these examinees were dropped for the purpose of the analyses. Table 2 also indicates that if the examinees with boundary estimates were included, the performance of the traditional CAT would have dropped precipitously.

Across both numbers of strata, the stratum CATs based on empirical difficulty estimates performed better than the conventional fixed length test. The sorting based on b-parameters and p-values had very similar squared correlations and essentially performed equivalently. The relative degree of adaptive benefit for the stratum CATs was calculated by comparing the increase in the squared correlation of the traditional CAT over the conventional test to the increase in the squared correlation of the stratum CAT over the conventional test. This index can be thought of as the percent of efficiency in the traditional CAT recovered by the stratum CAT. As the number of strata increased, the recovered efficiency in the stratum CAT increased. This pattern replicates Wise's (1999) findings and provides further evidence of the efficiency stratum CATs can provide when item assignment is completed using p-values.

The precision of the stratum CAT based on human judgment was worse than that of the conventional fixed length test for all conditions except the 20-item stratum CAT based on the pooled sortings across all judges from the Mathematics department. This stratum CAT only slightly outperformed the conventional test (recovered efficiency=7%). As expected, the stratum CATs based on human judgments, and all other tests, had greater precision as the test length increased. However, precision tended to decrease as the number of strata increased for the stratum CATs based on human judgments. This result can be explained by the decrease in the correlations between true and judged strata membership as the number of strata increase.

There was a benefit to pooling difficulty sortings for the 3-stratum CAT but not for the 5-stratum CAT. Specifically, for both groups of judges across both test lengths, the 3-stratum CAT based on judgments had the greatest precision when the pooled sortings were used. However, the 5-stratum CAT had the greatest precision when based on the best individual sorting.

These findings suggest that stratum CATs based on either group of human judges do not provide an increased efficiency over conventional fixed length tests. For all test conditions, the stratum CATs based on sortings from the Mathematics judges were slightly more precise than the stratum CATs based on sortings from the Educational Psychology judges. It appears in this case that expertise in the content area may have a greater benefit for sorting items than expertise in the area of test construction. However, again, none of the stratum CATs based on human judgments of item difficulty outperformed the conventional fixed-length test.

Algebra Items

The correlations between the true and judged item strata membership for the algebra items are presented in the bottom portion of Table 1. The values of the correlations are larger than those based on the ACT items and are more similar to those reported in previous studies (Lorge & Kurglov,1952, 1953; Quereshi & Fisher, 1977). Similar to the findings using the ACT items, correlations decreased as the number of item strata increased which shows the increased difficulty of making more precise judgments of relative item difficulty. The pooled judgments of item difficulty produced a higher correlation than the individual judgements for both the 5 and 7 strata sortings.

Table 3 presents the squared correlations between actual and estimated proficiency and the recovered efficiency of the stratum CATs. As was found with the ACT items, the stratum CAT based on empirical difficulty estimates had greater precision than the

conventional test and the recovered efficiency of the stratum CAT increased as the number of strata increased. The recovered efficiency of these stratum CATs decreased as test length increased, a finding opposite of that for the ACT items. Overall, the stratum CAT based on empirical difficulty estimates performed quite well. In addition to having greater precision than the conventional fixed length test, the stratum CAT had greater precision than the traditional CAT for test lengths of 15 and 20 items. While the relative efficiency decreased with increased test length, the index remained above 66% for the 5-strata test and above 80% for the 7-strata test. These findings provide further evidence of the increased precision and efficiency over conventional tests that is available from stratum CATs based on empirical difficulty estimates.

The stratum CATs based on human judgments had greater precision than the conventional tests only for 15 and 20-item test lengths. While there was one 5-stratum CAT based on an individual sorting that had greater precision than a conventional test for a 25-item test, this increase was slight (3%). Therefore, the following discussion will focus specifically on the 15 and 20-item test lengths. For both test lengths, the stratum CATs based on the pooled sortings from the judges had greater recovered efficiency than the stratum CATs based on individual sortings. Specifically, none of the stratum CATs based on individual sortings had greater precision than the conventional fixed-length test for the 20-item test. Only the pooled sortings produced a stratum CAT that would outperform the fixed-length test at this length. As was found with the ACT items, stratum CATs with fewer strata were relatively more efficient than stratum CATs with more strata. Specifically, the 5-stratum CAT had greater recovered efficiency than the 7-stratum CAT for both pooled and individual sortings. These findings provide evidence that the stratum CATs based on human judgments provide an increased efficiency over conventional fixed length tests when only a small number of items are

administered and provide maximum performance when items are sorted into fewer strata.

General Discussion

Two operational item pools were used to investigate the performance of stratum CATs when items were assigned to strata based on empirical estimates of item difficulty or human judgments of item difficulty. It was found that stratum CATs based on empirical item difficulties (both p-values and b-parameters) have increased efficiency and precision relative to a conventional fixed-length test. It was also shown that efficiency and precision increased as the number of strata increased. Also, under certain conditions the stratum CAT was able to match or exceed the precision and efficiency of the traditional CAT. These findings coupled with previous research (Wise 1999) provide evidence of the promise stratum CATs have as a non-IRT adaptive testing method that requires minimal item data.

The results from the stratum CAT based on human judgment were not promising. The findings when using an ACT mathematics item pool suggest that stratum CATs based on human judgments do not provide increased efficiency or precision over a conventional fixed-length test. Slightly more encouraging were the findings using an algebra item pool. It was found that stratum CATs based on human judgments outperformed a conventional fixed-length test when a small number of items were administered (15 or 20). In addition, the performance of the stratum CATs increased as the number of strata decreased and when item strata membership was based on pooled sortings across judges. This suggests that having individuals rank the difficulty of items for use in stratum CATs may only be justified if the test administration is very limited in terms of the number of items to be administered and empirical difficulty estimates are not readily available. If empirical estimates are available, a stratum CAT based on

these estimates would be more appropriate. If empirical estimates are not available but the test administration scenario allows for a longer test, the extra effort of having people rank items for use in a stratum CAT may not be advantageous over a traditional fixed length test. It is apparent from the results using both items pools that judges would have to sort items into strata substantially better than they performed in this study to benefit from the efficiency of the stratum CAT relative to a conventional fixed length test.

References

Green, K. E. (1983). Subjective judgment of multiple-choice item characteristics. Educational and Psychological Measurement, 43, 563-570.

Lorge, I. & Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. Educational and Psychological Measurement, 12, 554-561.

Lorge, I. & Kruglov, L. (1953). The improvement of estimates of test difficulty. Educational and Psychological Measurement, 13, 34-46.

Quereshi, M. & Fisher, T. L. (1977). Logical versus empirical estimates of item difficulty. Educational and Psychological Measurement, 37, 91-100.

Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A., & Severance, D. D. (in press). Examinee Judgments of Changes in Item Difficulty: Implications for Item Review in Computerized Adaptive Testing. Applied Measurement in Education.

Wise, S. L. (1999, April). Comparison of stratum scored and maximum likelihood-scored CATs. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Table 1

Correlations between true strata membership and judged strata membership

	Number of Strata		
Judges	3	5	7
ACT Mathematics item pool			
Ed Psych Pooled Judgements	.463	.343	
Ed Psych Rater A	.472	.468	
Ed Psych Rater B	.250	.193	
Ed Psych Rater C	.420	.200	
Ed Psych Rater D	.167	.200	
Math Pooled Judgements	.462	.358	
Math Rater A	.194	.167	
Math Rater B	.389	.301	
Math Rater C	.361	.258	
Math Rater D	.389	.325	
Algebra item pool			
Ed Psych Pooled Judgements		.588	.571
Ed Psych Rater E		.496	.481
Ed Psych Rater F		.550	.519
Ed Psych Rater G		.503	.500

Table 2

Squared correlations between true and estimated proficiency and recovered efficiency using the ACT Mathematics item pool

	Test Length	
Test	15	20
Conventional Fixed-Length Test	.780	.799
Traditional CAT	.832 .626[1]	.856 .658[2]
Stratum CATs (3 strata)		
Strata based on p-values	.787 (14)	.828 (51)
Strata based on b parameters	.790 (19)	.834 (61)
Ed Psych Pooled Judgements	.762 (0)	.796 (0)
Ed Psych Rater A	.750 (0)	.796 (0)
Ed Psych Rater B	.753 (0)	.785 (0)
Ed Psych Rater C	.743 (0)	.792 (0)
Ed Psych Rater D	.740 (0)	.776 (0)
Math Pooled Judgements	.769 (0)	.803 (7)
Math Rater A	.740 (0)	.773 (0)
Math Rater B	.746 (0)	.790 (0)
Math Rater C	.748 (0)	.797 (0)
Math Rater D	.753 (0)	.796 (0)
Stratum CATs (5 strata)		
Strata based on p-values	.801 (40)	
Strata based on b parameters	.797 (33)	
Ed Psych Pooled Judgements	.752 (0)	
Ed Psych Rater A	.767 (0)	
Ed Psych Rater B	.736 (0)	
Ed Psych Rater C	.733 (0)	
Ed Psych Rater D	.728 (0)	
Math Pooled Judgements	.769 (0)	
Math Rater A	.783 (0)	
Math Rater B	.760 (0)	
Math Rater C	.728 (0)	
Math Rater D	.746 (0)	

[1] squared correlation if 843 examinees with boundary MLE estimates of –5.0 or +5.0 were included.

[2] squared correlation if 790 examinees with boundary MLE estimates of –5.0 or +5.0 were included.

Note: The values in parentheses are the percent of recovered efficiency relative to the traditional CAT with boundary values excluded.

Table 3

Squared correlations between true and estimated proficiency and recovered efficiency using the algebra item pool

	Test Length			
Test	15	20	25	30
Conventional Fixed-Length Test	.701	.780	.817	.846
Traditional CAT	.773	.839	.878	904
Stratum CATs (5 strata)				
Strata based on b parameters	.799 (136)	.846 (112)	.870 (87)	.885 (67)
Ed Psych Combined Raters	.731 (42)	.790 (17)	.812 (0)	.834 (0)
Ed Psych Rater E	.719 (25)	.766 (0)	.805 (0)	.831 (0)
Ed Psych Rater F	.731 (42)	.778 (0)	.819 (3)	.843 (0)
Ed Psych Rater G	.707 (8)	.764 (0)	.808 (0)	.835 (0)
Stratum CATs (7 strata)				
Strata based on b parameters	.808 (148)	.852 (122)	.876 (97)	.893 (81)
Ed Psych Combined Raters	.729 (39)	.783 (5)	.808 (0)	.828 (0)
Ed Psych Rater E	.714 (18)	.774 (0)	.799 (0)	.828 (0)
Ed Psych Rater F	.712 (15)	.767 (0)	.806 (0)	.828 (0)
Ed Psych Rater G	.706 (7)	.760 (0)	.803 (0)	.826 (0)

Note. The values in parentheses are the percent of recovered efficiency relative to the traditional CAT.